CLAIMS

What is claimed is:

1. A method of processing queries in a network, comprising the steps of:

defining a plurality of virtual clusters from a plurality of servers; and

routing a request to a virtual cluster of the plurality of virtual clusters based on predetermined criteria in order to allocate system resources.

2. The method of claim 1, further comprising the steps of:

monitoring performance of the plurality of servers; and

sending a report in response to workload at one of the plurality of servers exceeding a pre-determined threshold so that routing of further requests to the one of the plurality of servers is altered.

3. The method of claim 2, further comprising the step of removing the one of the plurality of servers from an associated virtual cluster and adding the one of the plurality of servers back into the associated virtual cluster in response to workload falling below the pre-determined threshold.

4. The method of claim 2, wherein the sending a report sends a report to a network dispatcher and the network dispatcher performs the routing.

5. The method of claim 1, further comprising the steps of:

determining that one of the plurality of servers is overburdened based on statistics; and

reducing workload to the one of the plurality of servers if the statistics are above a threshold.

6. The method of claim 5, wherein the reducing step includes at least one of removing the one of a plurality of servers from one of the plurality of virtual clusters and limiting further requests from being routed to the one of a plurality of servers.

7. The method of claim 6, wherein the reducing step includes reassigning the one of a plurality of servers to another one of the plurality of virtual clusters.

8. The method of claim 1, wherein at least one of the plurality of servers is assigned to more than one of the plurality of virtual clusters.

9. The method of claim 1, wherein the predetermined criteria includes at least one of requester identity, requested application, time of day, day of week, and performance statistics.

10. The method of claim 9, wherein the requester identity is an internet address.

11. The method of claim 9, wherein the performance statistics include at least one of central processing unit (CPU) performance statistics, memory statistics, connection counts, throughput statistics, and response time statistics.

12. The method of claim 1, wherein the routing step includes selecting one of the plurality of virtual clusters for routing based on at least one of a requester's identity and a requested application.

13. The method of claim 12, further including selecting one server from the one of the plurality of virtual clusters for routing based on statistics.

14. The method of claim 13, wherein the selecting is based on performance statistics.

15. The method of claim 1, wherein at least one of the plurality of servers is at least one of a lightweight directory access protocol (LDAP) server and a web application server.

16. The method of claim 1, wherein the routing uses rules based routing.

17. The method of claim 1, further comprising the steps of reassigning one of the plurality of servers from one of the plurality of virtual clusters to another one of the plurality of virtual clusters, wherein the one of the plurality of virtual clusters has a workload below a threshold and the another one of the plurality of virtual clusters has a workload above the pre-determined threshold.

18. A method for load balancing servers, comprising the steps of:

allocating a plurality of servers among a plurality of virtual clusters;

monitoring the plurality of virtual clusters for workload capacity; and

reassigning at least one server from one of the plurality of virtual clusters to another of the plurality of virtual clusters based on workload capacity in order to reallocate system resources.

19. The method of claim 18, wherein the monitoring step includes determining when a workload capacity of the one of the plurality of virtual clusters has crossed a threshold based on statistics associated with the one of a plurality of virtual cluster's performance.

20. The method of claim 18, further comprising the step of identifying another of the plurality of virtual cluster having available workload capacity based on statistics associated with the virtual cluster's performance and transferring at least one of the plurality of servers to the another of the virtual cluster.

21. The method of claim 18, wherein the reassigning at least one server includes one of:

removing the server entirely from the one of a plurality of virtual cluster, and

assigning the at least one server to both the one of a plurality of virtual clusters and the another of the plurality of virtual clusters.

22. The method of claim 18, further comprising routing a request to one of the plurality of virtual clusters based on one of the requestor's identity, the requested application, and rules.

23. The method of claim 22, further comprising selecting one server assigned to the one of the plurality of virtual clusters based on statistics for routing the request.

24. A computer program product comprising a computer usable medium having readable program code embodied in the medium, the computer program product includes at least one component to:

define a plurality of virtual clusters from a plurality of servers; and

route a request to a virtual cluster of the plurality of virtual clusters based on predetermined criteria to allocate system resources.

25. The method of claim 24, wherein the at least one component:

monitors performance of the plurality of servers; and

sends a report in response to workload at one of the plurality of servers exceeding a pre-determined threshold so that routing of further requests to the one of the plurality of servers is altered.

26. The system of claim 25, wherein the at least one component removes the one of the plurality of servers from an associated virtual cluster and adding the one of the plurality of servers back into the associated virtual cluster in response to workload falling below the pre-determined threshold.

27. The system of claim 24, wherein the at least one component sends a report to a network dispatcher and the network dispatcher performs the routing.

28. The system of claim 24, wherein the at least one component:

determines that one of the plurality of servers is overburdened based on statistics; and

reduces workload to the one of a plurality of servers if the statistics are above a threshold.

29. The system of claim 28, wherein the at least one component removes the one of a plurality of servers from one of the plurality of virtual clusters and limits further requests from being routed to the one of a plurality of servers.

30. The system of claim 29, wherein the at least one component reassigns the one of a plurality of servers to another one of the plurality of virtual clusters to reallocate the system resources.

31. The system of claim 24, wherein the at least one component assigns at least one of the plurality of servers to more than one of the plurality of virtual clusters.

32. The system of claim 24, wherein the predetermined criteria includes at least one of requester identity, requested application, time of day, day of week, performance statistics.

33. The system of claim 32, wherein the requester identity is a network address.

34. The system of claim 32, wherein the performance statistics include at least one of central processing unit (CPU) performance statistics, memory statistics, connection counts, throughput statistics, and response time statistics.

35. The system of claim 24, wherein the at least one component selects one of the plurality of virtual clusters for routing based on at least one of a requester's identity, composite statistics, and a requested application.

36. The system of claim 24, wherein the at least one component selects a non over-burdened server from the one of the plurality of virtual clusters to process information.

37. The system of claim 36, wherein the at least one component selects based on performance statistics.

38. The system of claim 24, wherein at least one of the plurality of servers is one of a lightweight directory access protocol (LDAP) server and a web application server.

39. The system of claim 24, wherein the at least one component uses rules based routing.

40. The system of claim 24, wherein the at least one component reassigns one of the plurality of servers from one of the plurality of virtual clusters to another one of the plurality of virtual clusters, wherein the another of the plurality of virtual clusters has a workload below a threshold and the one of the plurality of virtual clusters has a workload above the pre-determined threshold.